The Causal Effect of Summer Job Experience on the Future Unemployment

Author: **Jianli Chu**
Supervisor: **Moudud Alam**

June, 2008.

Contents

Abstract

This paper examines the causal effect of early summer-job experience on the future unemployment. In order to get the causal effect of the treatment(summer job), we should exclude the self-select bias. The propensity score matching method is used in this study to correct the sample self-selection bias which causes the differences between the treatment and control groups. There are two kinds of respond variables in this study: the frequency of unemployment in one year and the duration of unemployment between two working periods. Multinomial model and Cox's regression model are applied to analyze the two outcomes respectively.
Keywords: Summer job, Unemployment, Matching, Propensity score, Cox's regression

1 Introduction

In the most European countries, the unemployment of youths keeps growing[1]. Experiences from the labour market of the developed courtiers, e.g. USA, Germany and Japan, reveal that the state of youth labour market, in general, chaotic at least for a substantial part of the youths (Gardecki and Neumark 1998). Referring to a 1994 report of the National Center for research in Vocational Training (NCREV) Gardecki and Neumark (1998)[5] further commented that the years spent by the youths in "floundering" from one job to another is a waste of human resources. The situation is likely to be the same in Sweden since Schröder (2004)[2] reported , on the basis of OECD statistics, that the youth (15-24 years) unemployment rate was 2.2 times the adult (25-54 years) unemployment rate during 1995-97. One of the causes for high youth unemployment rate may be the difficulty in getting the first job but regardless the cause a remedial suggested by Schröder (2004)[2] is to provide the school-leavers with work experience.

In fact, lots of Swedish students do summer job during June to July. Using Finish data Häkkinen (2004)[11] showed that in-school work experience increases the probability of being employed after graduation. Therefore, the aim of this work is to study the effect of summer-job on their future unemployment of the high school graduates in Sweden using a sample data from the whole Swedish data.

Since we want to analyze the causal effect of summer-job on the unemployment, we apply propensity score matching method to find the matched samples. Then two different outcomes about unemployment are considered as the respond variables, and two different models are set up to analyze the two respond variables. The results of the two models reveal that summer-job experience have no obvious causal effect on reducing the frequency of the unemployment and shortening the unemployment duration.

[1]According to Lena Schroder(2004), *The role of youth programmes in the transition from school to work*

2 Data Description

2.1 The Process of Original Data

The original data consists of 213,898 records which contain thousands of Swedish people's yearly information about working and individual's background. In this study, we defined summer-job as the treatment. If a person had worked during June to July between 16-19 years old, we can say he/she did summer job. Otherwise, he/she did not participate summer job.

The age of the people in this data ranges from 16 to 31. We suspect the summer-job has such a strong effect that its influence can remain after 10 years. Furthermore, the account of the last several years' records is much less than the early several years'. So we decide only consider the 20-25 information to analyze the unemployment situation.

We divide the whole data into two parts. One part is 16-19 subset data. The other one is 20-25 subset data. The 16-19 part is used to obtain matched samples by propensity score matching method. Then the people whose 16-19 information can be matched are considered, the samples cannot be matched on 16-19 part will be excluded in the further analysis.

It is possible that one person can experience more than one year with summer-job during 16-19 and this could be considered as increasing doses. But we do not consider the dose effect for about 81% people who did summer-job have no more than 2 summer-job experiences during 16-19. It means that the number of the individuals with more than 2 summer-job is very small, so maybe it can lead some bias result if considering the dose effect.

The individual's information was recorded by year. There is at least one record for one year. But, for example, if one individual has three different jobs in 2001, there are three records corresponding to the three jobs respectively for 2001. However, we would like to do a yearly data analysis and we are only interested in the unemployment situation of individuals rather than how many jobs he/she did in a year. Thus,

we have to merge the information from different records but within the same year and get a new record corresponding the year for the individual.

In order to pick up the important information we are interested and delete some useless information, the original data should be processed as below.
(a). We want to analyze whether the summer-job experiment effects the employment, so if there is no working information in the records we should delete them.
(b). Divide the data into two parts. One part is 16-19 subset data. The other one is 20-25 subset data.
(c). Delete the people whose information does not contain 16-19 years old information in 20-25 part and deleted the people whose information does not contain 20-25 information in 16-19 part. We cannot know the samples whether can be matched with others without 16-19 information, even if they have 20-25 information.
(d). Summarize the information and get the new information which is one-record-to-one-year form.
(e). For the individuals with summer job experience, we only considered the first year summer-job record and ignored the latter summer-job and non-summer-job records. This way can simplify multi-summer-job situation[2]. For the individuals without summer job experience, We keep all the records of non-summer-jobber.[3]

After a serial process, we obtain two sets of data. One part containing 16-19 years old information is used to find matched samples by applying propensity score matching method. The other one containing 20-25 years old information is used to analyze how the summer-job effects the unemployment. Since there are so many records that I have to use R to process this data set. The detail of procedure is available in Appendix.

[2]It is probably that one individual has more than one summer job records. The background information in these records may change. But we need the background information to *Matching*. Many variables of background information are categorical, which means it does not make sense if averaging them. Thus, it is hard to summarize the background information of summer job records for one individual.

[3]Since we have limited the records of summer-jobbers, a record cannot match with another one from the same individual. So keeping all the records of non-summer-jobbers can increase the probability of finding more matched samples.

2.2 The Description of Variables

There are four kinds of variables in every record of the data. The below table shows them.

Table 1 The list of variables used in the final analysis

	Description	Type of variables
Outcome		
Numb	the frequency of unemployment in one year	Discrete
Duration	the duration of unemployment	Discrete
Length	the length of unemployment in one year	Discrete
Background		
LopnrUt	ID number	Discrete
Ms-prob	Middle school grades(in percentile score)	Continuous
Kon	Gender(male=1,female=2)	Dummy
FodAr	Birthyear	Discrete
Lan	The person's living county,25 counties,(e.g. Stockholm,Dalarna,etc.)	Categorical 24 dummies
Kommun	The person's living municipality,339 municipalities,	Categorical 338 dummies
LoneInk	Wage earnings (in SEK)	Continuous
DispPersBasF	Per capita family income (in SEK)	Contiunuous
Ar	Current year	Discrete
Sektor	Sector where worked,101 sectors	Categorical 100dummies
Civi-lMor	Mother's marital status, 5 levels,(e.g. single,married,etc.)	Categorical 4 dummies
Sektor-Mor	Mother's working sector, 12 levels(e.g. Government administration,non-Socialist primary municipal administration etc.)	Categorical 11 dummies
SEI-Mor	Socio-economic status of mother, 14 levels(e.g. industrialist,farmer,etc.)	Categorical 13 dummies
Civil-Far	Father's marital status, as same as mother's marital status,	Categorical 4 dummies
Sektor-Far	Father's working sector, 13 levels(as same as mother's working sector)	Categorical 12 dummies
SEI-Far	Socio-economic status of father, 14 levels(as same as mother's SE status)	Categorical 13 dummies
Macro-economics		
GDP	The constant price index(year 2000=100)	Continuous
UErate	The unemployment rate of Sweden	Continuous
RU18-24	The 18-24 years old unemployment rate of Sweden in municipal level	Continuous
Treatment		
SJ	Summer jobber, 0=non-summer jobbers, 1=summer jobbers	Dummy

The outcome variables we are interested are the Frequency of Unemployment and Duration of Unemployment. These two variables are not given in the original data directly. The information of working we can know is which month with working and non-working in a year for every individual in the data. Therefore, we can summarize the information and get the variables we are interested. We defined the Frequency of Unemployment is the number of changes of employment to unemployment in one year, and the Duration of Unemployment is the number of continuous unemployed months.[4]

The background variables are used to find matching samples by propensity score matching method. The background variables include education, family, region, etc. They are given directly in the original data.

We add some macro-economics variables which have something with unemployment in the 20-25 part. They are the GDP(index, 2000 year as base year) of Sweden, the unemployment rate of Sweden both in national level and local level. The unemployment rate in national level is corresponding to the whole population in Sweden, while the unemployment rates in local level are corresponding to the every municipality. One point should be explained is that the municipal level unemployment rates are for the 18-24 years old population, which is used to approximately substitute for the unemployment rate for 20-25 year old in municipal level. Macro-economics variables are used to explain the outcome variables.

The treatment variable is SJ(summer job). It is the most important variable in this study.

3 Propensity Score Matching

3.1 Causal Effect

The first thing we should know is that we want to estimate the causal effect of treatment on the outcomes. Actually, there are two similar kinds of inferences, pre-

[4]The unemployed situation can keep going even to the next year. When this situation occurs, we define the duration belongs to the year when the duration starts.

dictive inference and causal inference. As Andrew and Jennifer(2007)[7] said,"In the usual regression context, predictive inference relates to comparisons between units, whereas causal inference addresses comparisons of different treatments if applied to the same units. More generally, causal inference can be viewed as a special case of prediction in which the goal is to predict what would have happened under different treatment options."

We define that the treatment group is the group participated summer-job and the control group is the group did not participate summer-job. Then the Average Treatment Effect for the treated population[5] is defined as

$$\tau \mid_{T=1} = E(\tau_i | T_i = 1) = E(Y_i^1 | T_i = 1) - E(Y_i^0 | T_i = 1) \tag{1}$$

Where $T_i = 1 (= 0)$ if the ith unit was assigned to treatment(control). Expected outcomes of the treatment, $E(Y_i^1 | T_i = 1)$, can be directly estimated by calculating the mean of Y_i^1 in the treatment group.

The first problem is there is no way to obtain the estimate of $E(Y_i^0 | T_i = 1)$. We never get the observations of $Y_i^0 | T_i = 1$. The nature idea for this problem is to substitute $E(Y_i^0 | T_i = 0)$ for $E(Y_i^0 | T_i = 1)$. Actually, we can make this replacement if randomization can be satisfied, because Randomization means:

$$Y_i^1, Y_i^0 \perp T_i \qquad E(Y_i^0 | T_i = 0) = E(Y_i^0 | T_i = 1) = E(Y_i | T_i = 0)$$

This means there is no systematically difference between treatment group and control group. So the treatment is not necessary, or the treatment can be *ignorable*.

The second problem is we cannot design an experiment which provides both treatment group and control group for the same unit, namely, $E(Y_i^0 | T_i = 0)$ and

[5]In a nonexperimental condition, The treatment effect for the treated group is different with the treatment effect for the untreated group. The reason is that the treatment and control samples are drawn from different subpopulation. In a randomized experiment,the treatment and control samples come from the same population, therefore, the treatment effects for the treatment and control groups are identical. Obvious, this case of summer-job is nonexperimental.

$E(Y_i^1|T_i = 1)$ are not available at the same time. In the other words, the same person has only one 16-19 period and cannot go back the past to make another choice. That means there is only one result of summer-job, did or not did. We never know what would happened if the same person made the other decision on summer-job.

3.2 Using Propensity Score Matching

Unfortunately, we cannot randomize this summer-job case like a experimental study. The students have the right to decide to do or not to do the summer-job and nobody can force them to be treated or not. Thus, the assumption of randomization cannot be satisfied in this study. This implies that it is probably that the treated samples and control samples come from different population. In other words, the people with different characteristics may make opposite choices on treatment. The differences between the characteristics may cause the distinct outcomes. For example, the students who took part in the summer job may have strong motivations to work, and also have a better employment situation than those who did not. This kind of situation is called self-selection bias. Thus, we have to consider the characteristics(X) of the individual.

For the second problem in the last section, if we can find the persons who have the same or similar characteristics(means no systematical difference) but make the different decisions, the problem can be solved in this way. In this case, we can pick up many pairs of students with same or similar characteristics, but one did summer-job while the other did not do summer-job in the each pair. Therefore, we need the information of characteristics of each individual to estimate the $\tau\ |_{T=1,X}$ first which can be done in the same way of $\tau\ |_{T=1}$:

$$\tau\ |_{T=1,X} = E(\tau_i|T_i = 1, X) = E(Y_i^1|T_i = 1, X) - E(Y_i^0|T_i = 0, X) \tag{2}$$

As the first problem in the last section shows, $E(Y_i^0|T_i = 1, X)$ is unobservable. We can take the weaker assumption that $Y_i^0 \perp T_i \mid \mathbf{X}$[6] to substitute $E(Y_i^0|T_i = 1, X)$ by

[6] $Y_i^0 \perp T_i \mid \mathbf{X}$ is enough to obtain the estimate of $\tau\ |_{T=1,X}$. The stronger assumption,$Y_i^1, Y_i^0 \perp T_i \mid \mathbf{X}$, need to identify the treatment effect on control group. In fact, we are interested in treatment effect on only treated group $\tau\ |_{T=1}$.

$E(Y_i^0|T_i = 0, X)$. After that, $\tau|_{T=1}$ can be obtained by averaging $\tau\ |_{T=1,X}$ over the distribution $X|T_i = 1$.

$$\tau|_{T=1} = E_X[(\tau|T = 1, X)|T_i = 1] \tag{3}$$

But how to pick up the students with same or similar characteristics? As Andrew and Jennifer(2007)[7] said *Matching*, "refers to a variety of procedures that restrict and reorganize the original sample in preparation for a statistical analysis". When there are quite few confounding variables, it is easy to match by some other matching method such as subclassification or Mahalanobis distance[7]. However, if there are many confounding covariates, the probability of finding the matched samples is getting smaller as the number of confounding covriates increasing. For example, if there are 10 different covariates in vector X, it is hard to find matched individuals with the 10 same values. we would better apply propensity score matching to create a one-number summary of all the confounding covariates and then use this number to match. Therefore, the multi-dimensional characteristics of every individual are simplified by only one number.

In fact, the propensity score is the probability of one person in the treatment group on the condition $X = x$:

$$Pr(T = 1|\mathbf{X}) = p(\mathbf{X}) = E(T|\mathbf{X}); 0 < p(\mathbf{X}) < 1 \tag{4}$$

The score can be estimated by regression models. In such models, treatment indictor is the outcome and all the confounding covariates are the predictors. Then the propensity scores can be calculated by using the estimated coefficient of the confounding covariates. The nature choice is setting up a logistic model:

$$logistic(P(x)) = \beta X + \epsilon \tag{5}$$

Transform this model and get the equation to calculate the score:

$$P(T = 1|\mathbf{X}) = p(\mathbf{X}) = \frac{1}{1 + exp(-\mathbf{X}\beta)} \tag{6}$$

[7]The detail can be seen in *Data analysis using regression and multilevel/hierarchical,pg.207*

Thus, we can substitute p(x) for the characteristics vector X and the treatment effect of treated group can be presented as below equation:

$$\tau|_{T=1} = E_{p(X)}[(\tau|T = 1, p(X))|T_i = 1] \tag{7}$$

The scores present the background information of individuals. When the scores are close, the corresponding individuals are considered as from the same population. Therefore, the self-selection bias and systematical difference do not exist.

4 Data Analysis

4.1 Getting the Matched Data

We have known the basic rules of propensity score matching, Then the process of calculating scores and matching them was carried out by R. The code of R is not given here, but it can be found on internet by the link in Appendix. The matching function we used is *Matchby*. We would like choose the option replacement[8] and one treated individual matched with two control individuals[9].

In the end, we obtain the matched data and the further data analysis is based on the matched data. Before the next analysis, we should check the result of matching.

Table 2 Some basic statistics about the summer jobbers and non-summer jobbers

		Summer Jobbers	Non-summer Jobbers	Total
Gender	male	0.33	0.67	1
		(1928)	(4036)	(5810)
	Female	0.30	0.70	1
		(1691)	(4036)	(5727)
Average middle school grade		0.57	0.55	0.56
	Total	3619	7918	15137

[8]Replacement means put the matched individuals back to the original data set after picking them up. Unreplacement works in a contrary way.

[9]This guarantee there are enough control samples matched the treated samples when the situation that one control sample match two treated samples happens.

Table 2 illustrates that the basic characteristics of the summer-jobber group and the Non-summer-jobber group seems very similar. Though there is a large gap between the size of the summer-jobber and Non-summer-jobber, we can accept this data for the highest difference of the proportion of the basic characteristics are not more than 3%. It shows that gender does not play a important role in choice of summer-job and summer-job does not depend on the grades of the students. The conclusion we can make is that there is no obvious *imbalance* and *lack of overlap*[10] on the Gender variable and ms-prob variable.

The Figure 1 prove the fact that matching procedure make the matched data more balance and overlap than before matching.

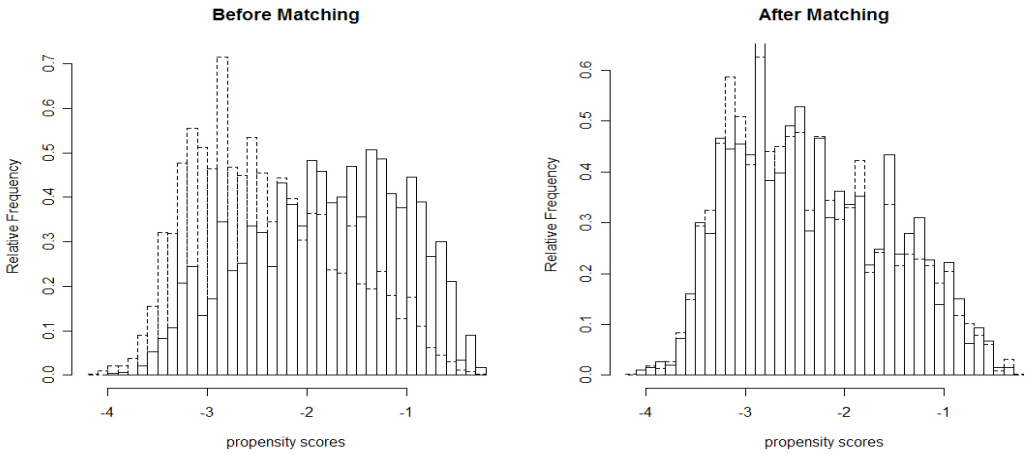

Figure 1 *(a) Histogram of propensity scores for treated(solid line) and control groups(dash line) before matching. (b) Histogram of propensity scores for treated(solid line) and control groups(dash line) after matching.*

[10]Imbalance occurs if the distributions of relevent pre-treatment variables differ for the treatment and control groups. Lack of complete overlap occurs if there are regions in the space of relevant pre-treatment variables where there are treated units but no controls, or controls but no treated units. More detail is available in *Data analysis using regression and multilevel/hierarchical*(Andrew and Jennifer,2007)pg.199-204

4.2 The Unemployment Frequency Model

After getting the matched data, we set up the model with the unemployment frequency as the respond variable. At the beginning, the poisson distribution models were considered. But almost all the poisson distribution models could not fit the data very well. Then we find that the multinomial models fit the data better. This can be proved by the below figure:

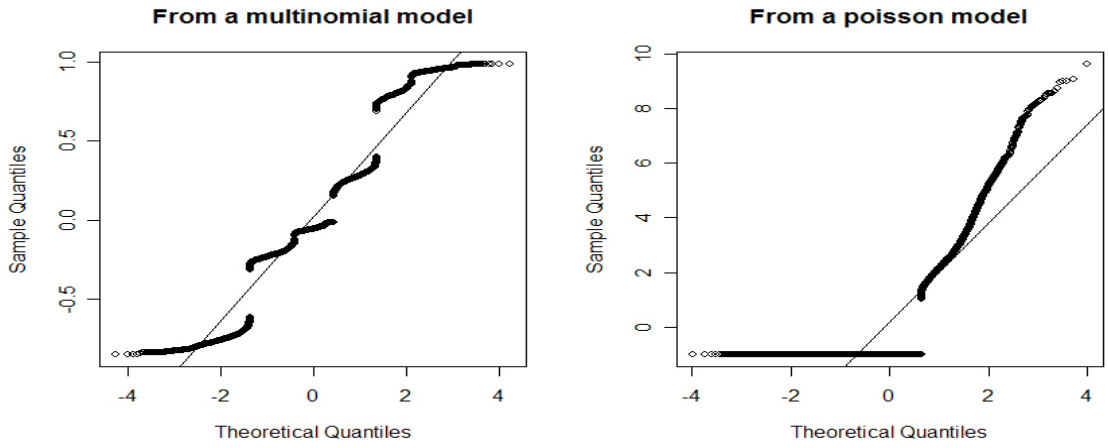

Figure 2 *(a) The qq-plot of a multinomial model (b) qq-plot of a poisson model*

Thus, it is reasonable to find the proper There are four levels of the unemployment frequency, 0,1,2,3. We discover that there are fewer observations with 3 times unemployment comparing with the observations with 1 and 2 times unemployment. So we decide to merge the 2 times level and the 3 times level. Therefore, there are left 3 levels of unemployment frequency, and this change improve the fitting result. The new variable "NN" substitute for the four-levels variable "Numb" in the model. However, the multinomial models display the results out of what we expected. Here is a typical nultinomial model:

Call: multinom(formula = factor(NN) ~ factor(Age) + SJ + sqrt(RU), data = addnew)

	coefficinets1	SE1	coefficients2	SE2
intercept	-1.625028	0.08978617	-4.468867	0.21172376
SJ	0.1657413	0.05318423	0.2302370	0.09873258
sqrt(RU)	0.1473946	0.02626687	0.1843005	0.04949904
factor(Age)21	0.08262168	0.0619231	1.55955746	0.1665862
factor(Age)22	0.02355861	0.06372928	1.50409462	0.16869125
factor(Age)23	-0.1138851	0.06674152	1.4645906	0.17072848
factor(Age)24	-0.125865	0.06871208	1.188440	0.17916142
factor(Age)25	-0.4307451	0.07550577	0.9139779	0.18921700
Residual Deviance: 21436.98		AIC: 21468.98		

From the above result of the model, we can see the trend of the coefficients of Age is decrease, which means the number of unemployment frequency is decreasing as the Age becoming larger. The coefficients of variable RU are positive which means that the higher youth unemployment rate in municipal level cause the more unemployment frequency. We also notice that the coefficients of SJ are positive, which means summer-job experience cannot reduce unemployment frequency and even maybe increase the times of unemployment. It is hard to explain. However, we have to accept this fact because we have tested almost all the possible models.

4.3 The Unemployment Duration Model

First, we would like to briefly introduce something regarding *Cox's proportional hazards model*(*Cox's regression* for short). This kind of model is quite commonly used in medical statistics study where the main outcome variable is the time to the occurrence of a particular event, for instance, the survival time until a patient dead. In reality, there are also many other response variables are the time to a certain endpoint. In this case, we have a similar outcome variable-*duration* which is the time from a unemployment to a employment occurs. Hence, we would rather using Cox's regression than the others regression. According to Brian and Torsten(2006)[3] said,Such data generally need special method to analysis for two mainly reasons:
(a). Survival data are generally not symmetrically distributed-they will often appear positively skewed, with a few people surviving a very long time compared with the majority; so assuming a normal distribution will not be reasonable.
(b). At the completion of the study, some patients may not have reached the endpoint of interest(death,relapse,etc.). Consequently, the exact survival times are not known.

All that is known is that the survival times are greater than the amount of time the individual has been in the study. The survival times of these individual are said to be *censored*(precisely, they are right-censored).

There are two function, namely the *survival function* and the *hazard function* play important roles in Cox's regression.

4.3.1 The Survival Function

The survival function, $S(t)$, is defined as the probability that the survival time, T, is greater than or equal to some time t, i.e.,

$$S(t) = P(T \geq t) \tag{8}$$

If some censored individuals are in the sample of survival times, the *Kaplan-Meier* estimator is required to estimate survival function. This involves first sorting the survival times from the smallest to the largest such that $t_{(1)} \leq t_{(2)} \leq ... \leq t_{(n)}$. The Kaplan-Meier estimate of the survival function is obtained as

$$\hat{S}(t) = \prod_{j:t(j)\leq t} (1 - \frac{d_j}{r_j}) \tag{9}$$

where r_j is the number of individuals at risk just before $t_{(j)}$ (including those censored at $t_(j)$), and d_j is the number of individuals who experience the event of interest(unemployment in this study) at time $t_{(j)}$. Thus, the survival function at the second unemployment time, $t_{(2)}$ is equal to the estimated probability of unemployment at time $t_{(2)}$, conditioning on the individuals have a positive probability to be employed at time $t_{(2)}$.
The estimated variance of the Kaplan-Meier estimate of the survival function is found from

$$Var(\hat{S}(t)) = (\hat{S}(t))^2 \sum_{j:t(j)\leq t} \frac{d_j}{r_j(r_j - d_j)} \tag{10}$$

4.3.2 The Hazard Function

The hazard function, $h(t)$, is defined as the probability that an individual experiences the event in a small time interval, when the size of the time interval approaches zero. It can be written as

$$h(t) = \lim_{s \to 0} P(t \leq T \leq t + s | T \geq t) \tag{11}$$

The relationship between the survival function and the hazard function is

$$S(t) = exp(-H(t)) \tag{12}$$

where $H(t)$ is known as the integrated hazard or cumulative hazard, and is defined as $H(t) = \int_0^t h(u)du$.
The hazard function can be estimated as the proportion of individuals experiencing the event of interest in an interval per unit time, given that they have survived to the beginning of the interval, that is

$$\hat{h}(t) = \frac{d_j}{n_j(t_{(j+1)} - t_{(j)})} \tag{13}$$

The estimate of $H(t)$

$$\hat{H}(t) = \sum_j \frac{d_j}{n_j} \tag{14}$$

4.3.3 Cox's Regression

The most popular form of Cox's regression is

$$log(h(t)) = log(h_0(t)) + \beta_1 x_1 + ... + \beta_q x_q \tag{15}$$

where $h_0(t)$ is known as the baseline hazard function, being the hazard function for individuals with all explanatory variables equal to zero. The model can be rewritten as

$$h(t) = h_0(t) exp(\beta_1 x_1 + ... + \beta_q x_q). \tag{16}$$

In this model, the base line hazard describes the common shape of the survival time distribution for all individuals, while the relative risk function, $exp(\beta_j)$ gives the relative risk change associated with an increase of one unit in covariate x_j, all other explanatory variables remaining constant. The result of Cox's regression seems easier to explain than the multinomial model with unemployment frequency.
Now, we have known the basic regulations of Cox's regression. The model can be set up by using R.

Call: coxph(formula = Surv(duration, event) SJ + factor(Age) + frailty(LopnrUt, distribution = "gamma"), data = addnew) n= 16414

	coef	se(coef)	se2	Chisq	DF	p
SJ	-0.064	0.0267	0.0222	5.75	1	1.7e-02
factor(Age)21	0.171	0.0256	0.0251	44.59	1	2.4e-11
factor(Age)22	0.220	0.0263	0.0257	70.23	1	0.0e+00
factor(Age)23	0.294	0.0271	0.0264	118.35	1	0.0e+00
factor(Age)24	0.365	0.0279	0.0272	172.09	1	0.0e+00
factor(Age)25	0.321	0.0299	0.0292	115.52	1	0.0e+00
frailty(LopnrUt, distribu				1616.23	1038	0.0e+00

	exp(coef)	exp(-coef)	lower .95	upper .95
SJ	0.938	1.066	0.89	0.988
factor(Age)21	1.186	0.843	1.13	1.247
factor(Age)22	1.246	0.802	1.18	1.312
factor(Age)23	1.342	0.745	1.27	1.415
factor(Age)24	1.441	0.694	1.36	1.522
factor(Age)25	1.379	0.725	1.30	1.462

Iterations: 9 outer, 39 Newton-Raphson
Variance of random effect= 0.0928 I-likelihood = -140900.2
Degrees of freedom for terms= 0.7 4.8 1038.2
Rsquare= 0.154 (max possible= 1)
Likelihood ratio test= 2743 on 1044 df, p=0
Wald test = 242 on 1044 df, p=1

From the above table, we can see the coefficient of SJ variable is negative. This means when a student did summer job, it is probably he or she has a smaller probability of ending the duration of unemployment between 20 and 25 years old. It is easy to know by the term of expcoef in the table. This term means the ration of $\frac{h(t|x+1)}{h(t|x)}$.

Actually, we have known the model form, that is

$$log(h(t)) = log(h_0(t)) + \beta_1 x_1 + ... + \beta_q x_q \tag{17}$$

And it can be rewritten as

$$h(t) = h_0(t) exp(\beta_1 x_1 + ... + \beta_q x_q). \tag{18}$$

Thus, the term of exp(coef) is presented as

$$\frac{h(t|x+1)}{h(t|x)} = \frac{exp(\beta^T(X+1))}{exp(\beta^T X)} = exp(\beta) \tag{19}$$

The exp(coef) of SJ is 0.94, which indicates $\frac{h(t|x+1)}{h(t|x)} = 0.94$ and reveals summer-jobbers have smaller probability to end the unemployment duration than the non-summer-jobbers. In other words, summer-job experience cannot help the students get a job in a shorter time.

The term of $frailty(LopnrUt, distribution = "gamma")$ means excluding the individual effect on the respond variable. For example, one student always have a long duration of unemployment and it is probable that he still have a long duration in the future. The function of this term is to delete this kind of relationship. It turns out that the result with the term is better than without it, for the Rsquare is improved.

5 Conclusion

The goal of this paper is to study the causal effect of summer-job experience on the future unemployment, using a sample data from the whole of Sweden. Since the student have right to make the choice of whether participate summer job and this study is non-experimental, we have to consider the self-selection bias in a causal effect study. In order to exclude the self-selection bias, we apply the propensity score matching method to cancel out this kind of bias. After matching procedure, we obtain the data without self-selection bias. Then the data is analyzed in two ways, one focus on the unemployment frequency while the other one focus on the unemployment duration. These two way have a common conclusion that summer-job experience has

no obvious effect on reducing unemployment frequency or shorten the unemployment duration.

However, we cannot say that summer-job experiences have no help for the future labor market perform. There are many standards to evaluate the perform in the future labor market except the standard of unemployment. A relative study made a conclusion that "the summer-job experience has a significant long term effect in the future earnings of the high school students",said by Wang and Alam(2007)[6]. In my opinion, summer job experience may help a person with the future labor market perform, but this paper illustrates summer-job experience in high school has no obvious effect on reducing unemployment frequency and shortening the unemployment duration of the youth(20-25 years old) in Sweden.

6 References

[1]Rajeev H.Dehejia and Sadek Wahba(2002), *Propensity score-matching methods for nonexperimental causal studies*,
http://www.personal.ceu.hu/departs/personal/Gabor_Kezdi/Program-Evaluation/Dehejia-Wahba-2002-matching.pdf

[2]Lena Schröder(2004), *The role of youth programmes in the transition from school to work*, IFAU- Institute for labour market policy evaluation(ISSN 1651-1166).

[3]Brian S. Everitt and Torsten Hothorn(2006), *A handbook of statistical analyses using R*, Boca Raton chapman & Hall/CRC.

[4]Box-Steffensmeier. Janet M.(2004), *Event history modeling: A guide for social scientists*, Cambridge University Press.

[5]Rosella Gardecki and David Neumark(1998), *Order from chaos?The effects of early labor market experiences on adult labor market outcomes*, Industrial and labor relations review, Vol.51,No.2.pp.299-322.

[6]Iris J.Y. Wang and Moudud Alam(2007), *When are non-experimental estimates close to experimental estimates? Evidence from a study of summer job effects in Sweden*, Uppsala universitet, Uppsala

[7]Andrew Gelman and Jennifer Hill(2007), *Data analysis using regression and multilevel/hierarchical*, Cambridge university press.

[8]Moeschberger. Melvin L.(1997), *Survival analysis*, Springer−Verlag New York.

[9]Laura Larsson(2003), *Evaluation of Swedish youth labor market programs*, The journal of huaman resources, Vol.38,No.4.PP.891-927.

[10]Tony Lancaster(1979), *Econometric methods for the duration of unemployment*, Econometrica,Vol.47,No.4.pp.939-956

[11]Iida Häkkinen(2004), *Working while enrolled in a university: does it pay?*, http://www.nek.uu.se/pdf/wp2004_1.pdf

7 Appendix

The R code of this thesis can be found by clicking this link:
http://user.qzone.qq.com/356343555/blog/1213109314

The data of unemployment is from:
http://mstatkommun.arbetsformedlingen.se/default.aspx?p=hist

The data of GDP is from:
http://www.scb.se/Statistik/NR/NR0103/2008K01A/NR0103
_2008K01A_DI_01_EN_BNP1950.xls

If the original data is required, please contact me. (h07jiach@du.se)